

COMPSCI 674 - Intelligent Visual Computing

DreamConstructorPro

Optimizing Meshes to align with Textual Prompts

- Vara Prasad Gudi
- Amritansh Mishra
- Manas Madine
- Priya Yarrabolu

Guided by : Evangelos Kalogerakis

AGENDA



- Introduction
- Related Works
- Model And Architecture
- Experiments Performed
- Results and Analysis
- Future Scope
- Conclusion

INTRODUCTION

- Traditional holistic approaches for text-to-3D scene generation have limitations in **capturing complex scenes with multiple objects** and intricate relationships using vectorized text embeddings.
- The **core** of our pipeline involves three stages :
 - Generating **individual mesh** objects using the image-to-3D pipeline from *DreamGaussian*.
 - Extracting **initial mesh configurations** from large language models (e.g., *GPT4*).
 - **Optimizing** the positions, orientations, and scales of these meshes based on two approaches: **CLIP scores and SDS loss**.

RELATED WORKS

- Ideas and approaches inspired from:
 - Optimization - Based 2D Lifting Methods
 - 3D Native Generative Models
 - Similar approaches and baselines

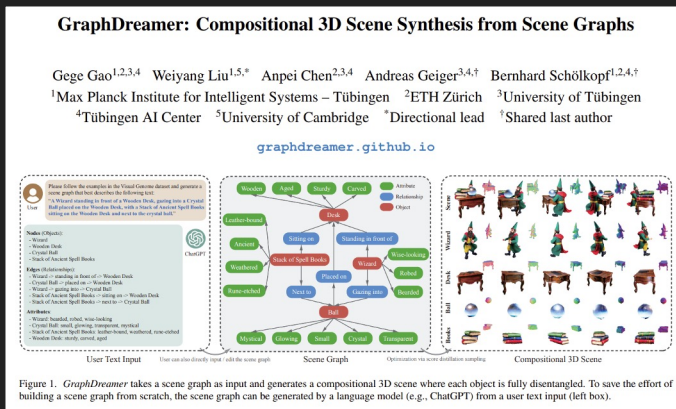


Figure 1. *GraphDreamer* takes a scene graph as input and generates a compositional 3D scene where each object is fully disentangled. To save the effort of building a scene graph from scratch, the scene graph can be generated by a language model (e.g., ChatGPT) from a user text input (left box).

<https://arxiv.org/pdf/2312.00093>

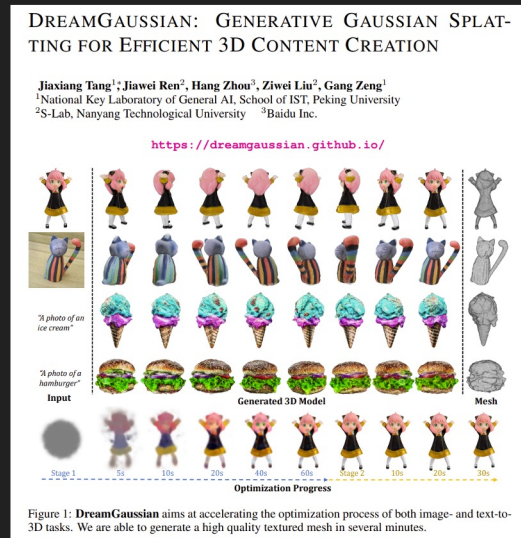


Figure 1: **DreamGaussian** aims at accelerating the optimization process of both image- and text-to-3D tasks. We are able to generate a high quality textured mesh in several minutes.

<https://arxiv.org/pdf/2309.16653>

RELATED WORKS

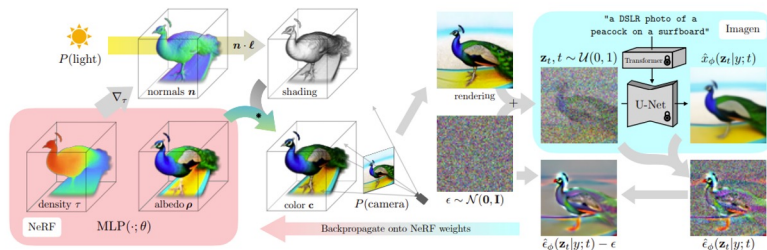


Figure 3: DreamFusion generates 3D objects from a natural language caption such as “a DSLR photo of a peacock on a surfboard.” The scene is represented by a Neural Radiance Field that is randomly initialized and trained from scratch for each caption. Our NeRF parameterizes volumetric density and albedo (color) with an MLP. We render the NeRF from a random camera, using normals computed from gradients of the density to shade the scene with a random lighting direction. Shading reveals geometric details that are ambiguous from a single viewpoint. To compute parameter updates, DreamFusion diffuses the rendering and reconstructs it with a (frozen) conditional Imagen model to predict the injected noise $\hat{\epsilon}_\phi(z_t|y; t)$. This contains structure that should improve fidelity, but is high variance. Subtracting the injected noise produces a low variance update direction $\text{stopgrad}[\hat{\epsilon}_\phi - \epsilon]$ that is backpropagated through the rendering process to update the NeRF MLP parameters.

<https://arxiv.org/pdf/2209.14988>

Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation

Haochen Wang^{*1} Xiaodan Du^{*1} Jiahao Li^{*1} Raymond A. Yeh² Greg Shakhnarovich¹
¹TTI-Chicago ²Purdue University

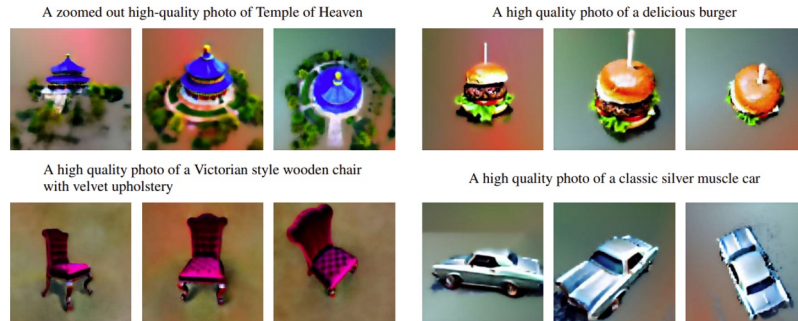
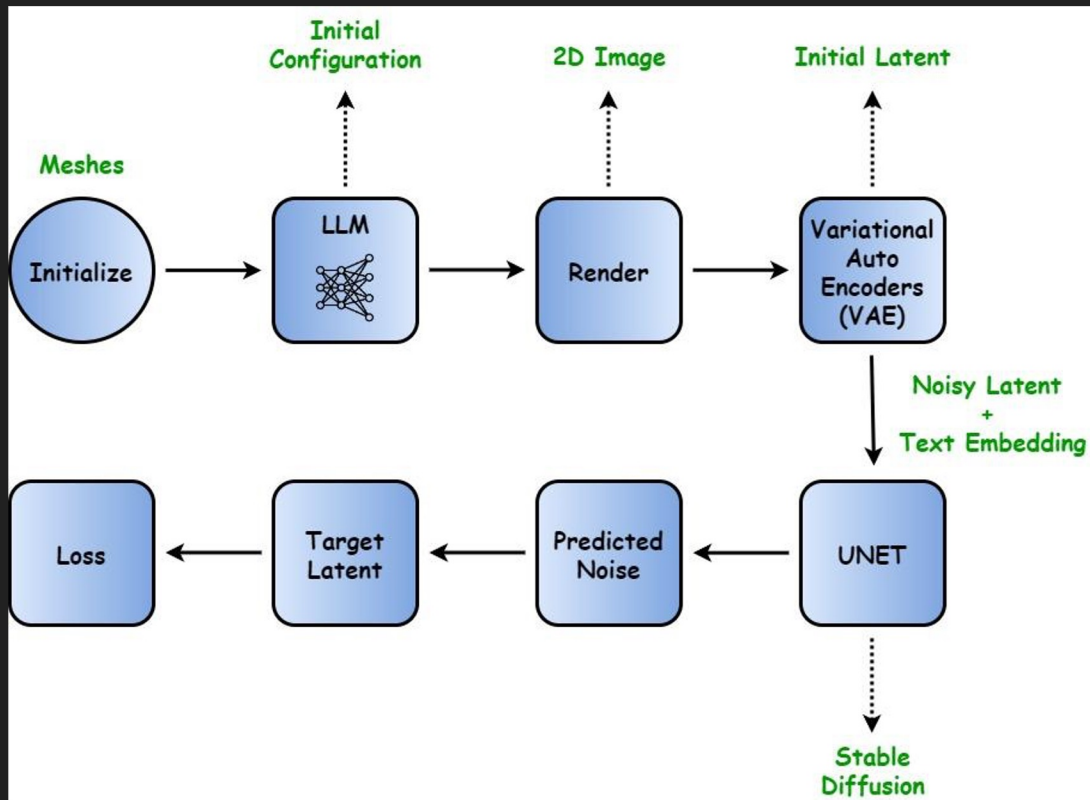


Figure 1. Results for text-driven 3D generation using Score Jacobian Chaining with Stable Diffusion as the pretrained model.

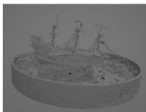
2D Diffusion Models For 3D Generation

Also, credits to: CompoDreamer (<https://github.com/justin871030/CompoDreamer>)

MODEL AND ARCHITECTURE



EXPERIMENTS PERFORMED

	PROMPT	OBJECTS	INITIAL 3D SCENE	OPTIMIZED 3D SCENE		ITERATION / EPOCHS
				CLIP LOSS	SDS LOSS	
1.	A CHAIR AND A TABLE WITH A TOY DINOSAUR ON IT					3500
2.	A TOY DINOSAUR ON TABLE					2400
3.	A SHIP IN THE GARDEN					3200
4.	A HOTDOG ON THE CHAIR					2599
5.	A TOASTER AND A HOTDOG ON THE TABLE					1800

RESULTS & ANALYSIS

- Refinement over 3500 iterations
- Optimizing the spatial arrangement, orientation, and scaling of the chair, table, and toy dinosaur to achieve a coherent and accurate representation of the described scene aligning to the input prompt.
- Source code to our work:

<https://github.com/amritansh6/DreamConstructorPro>



FUTURE SCOPE

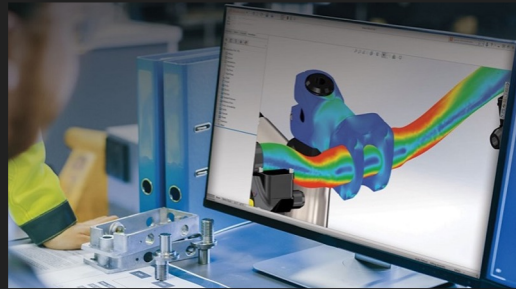
- A novel Diffusion-GAN dual training strategy to guide the training of 3D models.
- Efficiently generate 3D Gaussian representations from text input using Structured Volumetric Representation, Coarse-to-fine Generation Pipeline, etc.
- Using a text-conditioned tri-plane latent diffusion model and then use 2D diffusion priors to further refine the texture of coarse 3D models.
- Model the 3D parameter as a random variable instead of a constant as in SDS using Variational Score Distillation.
- Combine multiple noise estimation processes with a pretrained diffusion prior.
- Approach to use to learn and improve text-to-3D models from human preference feedback.
- Explicit injection of 3D priors from retrieved reference objects without re-training the 2D diffusion model.

REAL-TIME APPLICATIONS

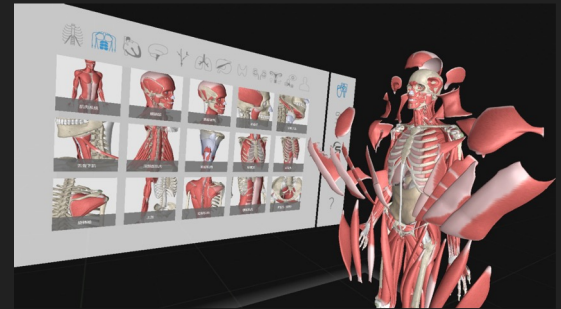
- Enhanced Scene Realism directly influenced by textual prompts.
- **Expansion to Interactive Environments:**
 - Extend capabilities to create interactive 3D environments that can respond to user inputs in real-time.



Construction visualization



Simulation Tools



Virtual Anatomy - AR/VR combination

CONCLUSION

- Bridging the gap between NLP and 3D scene generation, offering a trade-off between generation **quality** and **computational** efficiency.
- Enables precise grounding of textual entities and facilitates the creation of detailed scenes.
- Integrated with scene optimization and data augmentation, it could streamline large-scale text-to-3D pipelines for various content creation domains.
- Struggles with highly complex scenes and is dependent on the performance of underlying components like the image-to-3D pipeline and language models.

THANK YOU

